Proposed Approach for Modeling Fuel Effects on Air Toxics

EPA/NVFEL

EPAct / V2 / E-89 Program Review

Ann Arbor, MI

August 24, 2010

Acknowledgements

- Adam Sales (EPA intern)
- Bob Mason (SwRI)
- Kevin Whitney, Chris Sharp (SwRI)
- Dick Gunst (Southern Methodist Univ.)
- Aron Butler, David Hawkins, Cay Yanca, Michael Christianson (EPA)

Study Parameters for Selected Toxics

	Bag 1	Bag 1	Bags 2,3
Fuels	**27**	**11**	**11**
Vehicles	**15**	**5**	**5**
Replicates	**2 per run**	**none**	**none**
Compounds	acetaldehyde formaldehyde acrolein ethanol	benzene 1,3 butadiene	acetaldehyde formaldehyde acrolein ethanol benzene 1,3 butadiene
Fixed Model	ethanol RVP aromatics T50 T90 etOH*etOH T50*T50 RVP*etOH arom*etOH T50*etOH T90*etOH	ethanol aromatics T50 T90	ethanol aromatics T50 T90
Random Model	Vehicle	Vehicle	Vehicle
Censoring	yes?	**YES**	**YES**

Reduced Fuel Matrix
(*n* = 11)

Highlighted fuels, (except fuel 4), Used for Bag-1 (HALF) and Bag-2,3 analyses

Fuel	etOH	RVP	arom	T50	T90
1	10	10	15	150	300
2	0	10	15	240	340
3	10	7	15	220	300
4	10	**10**	15	220	340
5	0	7	35	240	300
6	10	7	15	190	340
7	0	7	15	190	300
8	0	10	15	220	300
9	0	10	35	190	340
10	10	7	35	220	340
11	10	10	35	190	300
12	10	10	35	150	340
13	0	7	35	220	340
14	0	7	15	190	340
15	0	10	35	190	300
16	10	7	35	220	300
20	20	7	15	165	300
21	20	7	35	165	300
22	20	10	15	165	300
23	20	7	15	165	340
24	20	10	15	165	340
25	20	10	35	165	340
26	15	10	35	165	340
27	15	7	15	220	340
28	15	7	35	220	300
30	10	10	35	150	325
31	20	7	35	165	325

Designing the (Full) Fuel Matrix
(for Phase 3)

- Fuel matrix based on computer-generated "optimal design"
  - Need to reduce test runs
  - Fuel properties correlated
- In "optimal design"
  - Fuel properties "*nearly orthogonal*"
  - Estimated effects (β's) correlated (somewhat)
- In contrast to standard factorial design, in which
  - Factors would be orthogonal
  - Estimated effects uncorrelated (independent)

Evaluating the Matrix

- Optimal design evaluated in terms of “efficiency”
  - Indicates how design approximates orthogonal factorial
  - Standard factorial 100% efficient
- Efficiency is function of
  - Number of fuel properties
  - Number of test points
  - Effects to be estimated (main, interactions)
  - “max std error for prediction” over the design points
- Criterion: efficiency > 50% considered “good enough”

Reevaluating the Reduced Matrix

- Design efficiency initially reviewed for full matrix
  - By Bob Mason, SwRI
- Reduced matrix represents an effective design change
  - review of efficiency needed
  - Question: what effects can be estimated?

Results

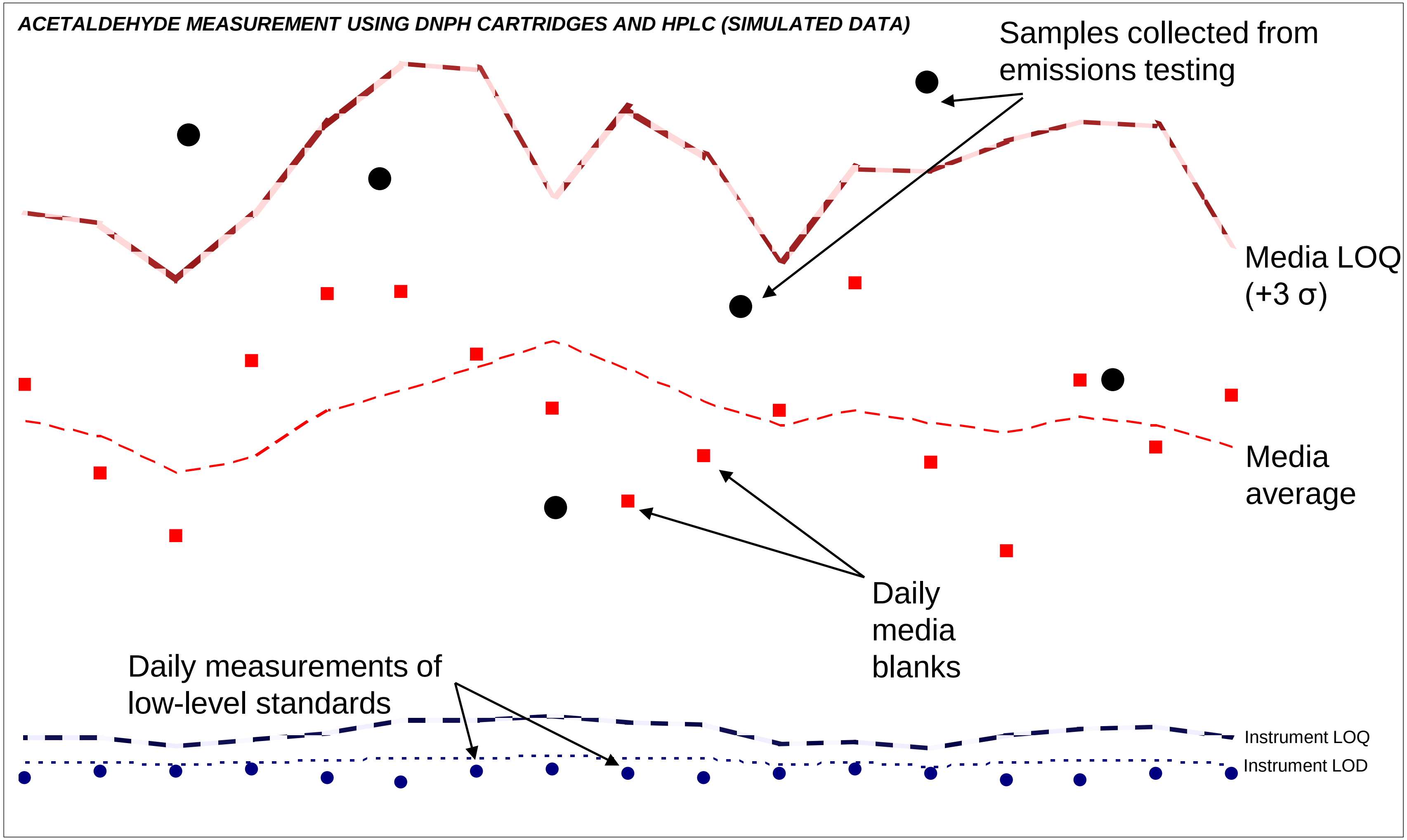
	Design	test fuels	Model terms	G-efficiency
	Full	27	ALL	51.6
Main Effects	1a	12	etOH, RVP, ARO, T50 T90	15.1
Main Effects	1b	12	etOH, ARO, T50 T90	48.8
Main Effects	**1c**	**11**	**etOH, ARO, T50 T90**	**58.3**
Interactions	2	11	1c + etOH*ARO	21.1
Interactions	3	11	2 + etOH*etOH	17.1
Interactions	4	11	3 + etOH*T50	2.8
Interactions	5	11	4 + etOH*T90	3.0

... And the winner is ... design 1c

NOTE: G-efficiency is expressed in relation to a hypothetical orthogonal design that cannot be realized. It is best viewed as a relative ranking among designs, rather than an absolute measure.

“Censoring”

- Having measurements recorded as
  - Non-detect, or
  - Below reporting limits
    - Limit of quantitation (LOQ): level at which we are confident that we have a meaningful quantitative value.
  - Affecting “lower tail” or “left-side” of distribution
  - Data “multiply censored” in reporting limits variable
- Common issue in environmental field
  - When measuring contaminants in
    - Water, soil, sediment, tissue, air, etc.


ACETALDEHYDE MEASUREMENT USING DNPH CARTRIDGES AND HPLC (SIMULATED DATA)
Samples collected from emissions testing
Media LOQ (+3 σ)
Media average
Daily media blanks
Daily measurements of low-level standards
Instrument LOQ
Instrument LOD

Some Definitions

- Instrument LOD shown here = 5-day running average of low level standard + (3 x std dev of 5-day running set of low level standards)

- Instrument LOQ shown here = 5-day running average of low level standard + (10 x std dev of 5-day running set of low level standards)

- Media average shown here = 5-day running average of media blanks

- **Media LOQ** shown here = 5-day running average of media blanks + (3 x std dev of 5-day running set of media blanks)

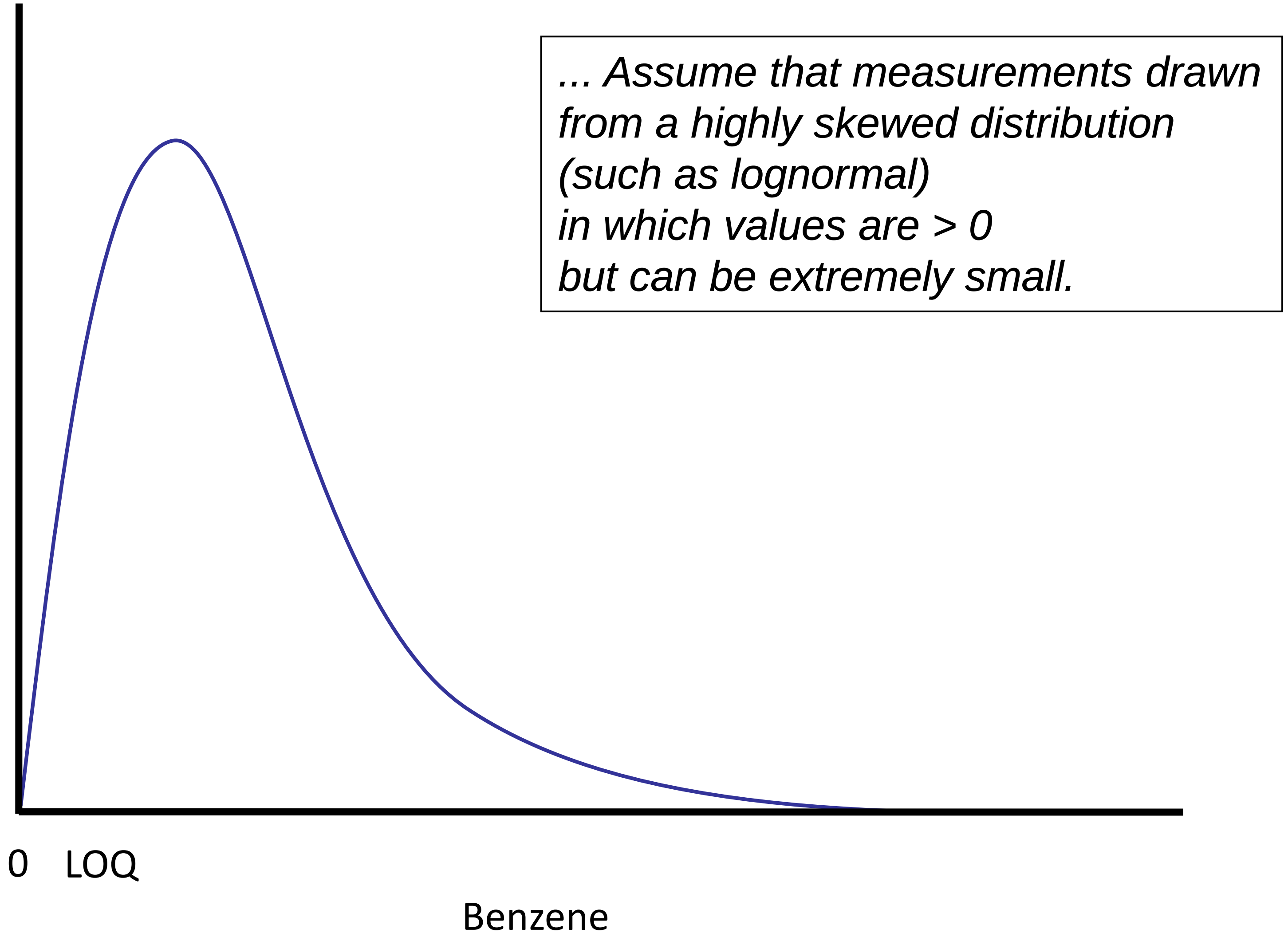
- Relative levels of instrument and media averages were taken from actual data

Left Censoring:
censoring rates (%)

	Bag 1	Bag 2	Bag 3
acetaldehyde	0	1.4	61
formaldehyde	0	1.4	1.4
acrolein	14	95	100
ethanol	23	42	78
benzene	0	69	80
1,3 butadiene	0.5	66	93

Missing = "below limit of detection" (<LOD)
OR "below limit of quantitation" (< LOQ)

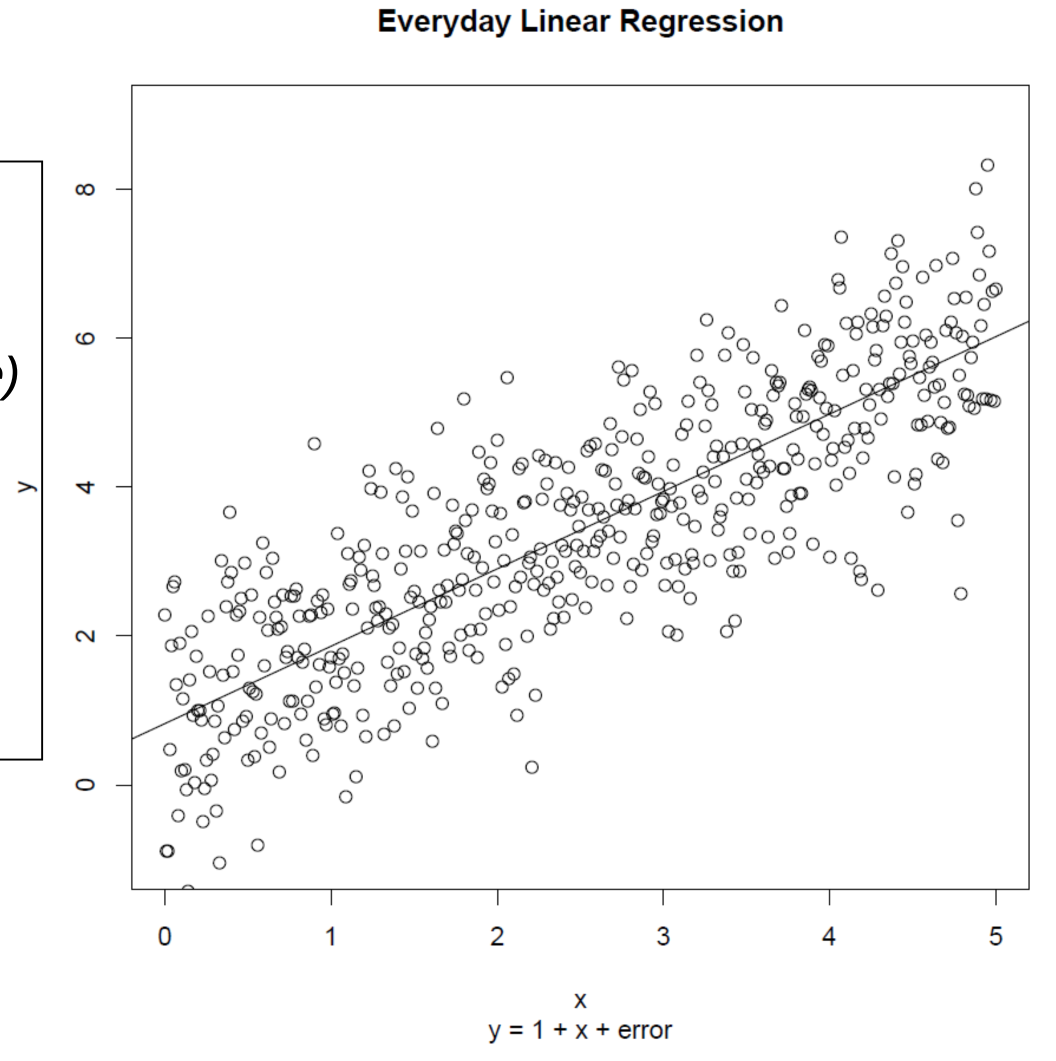
Uncensored Distribution



An Example

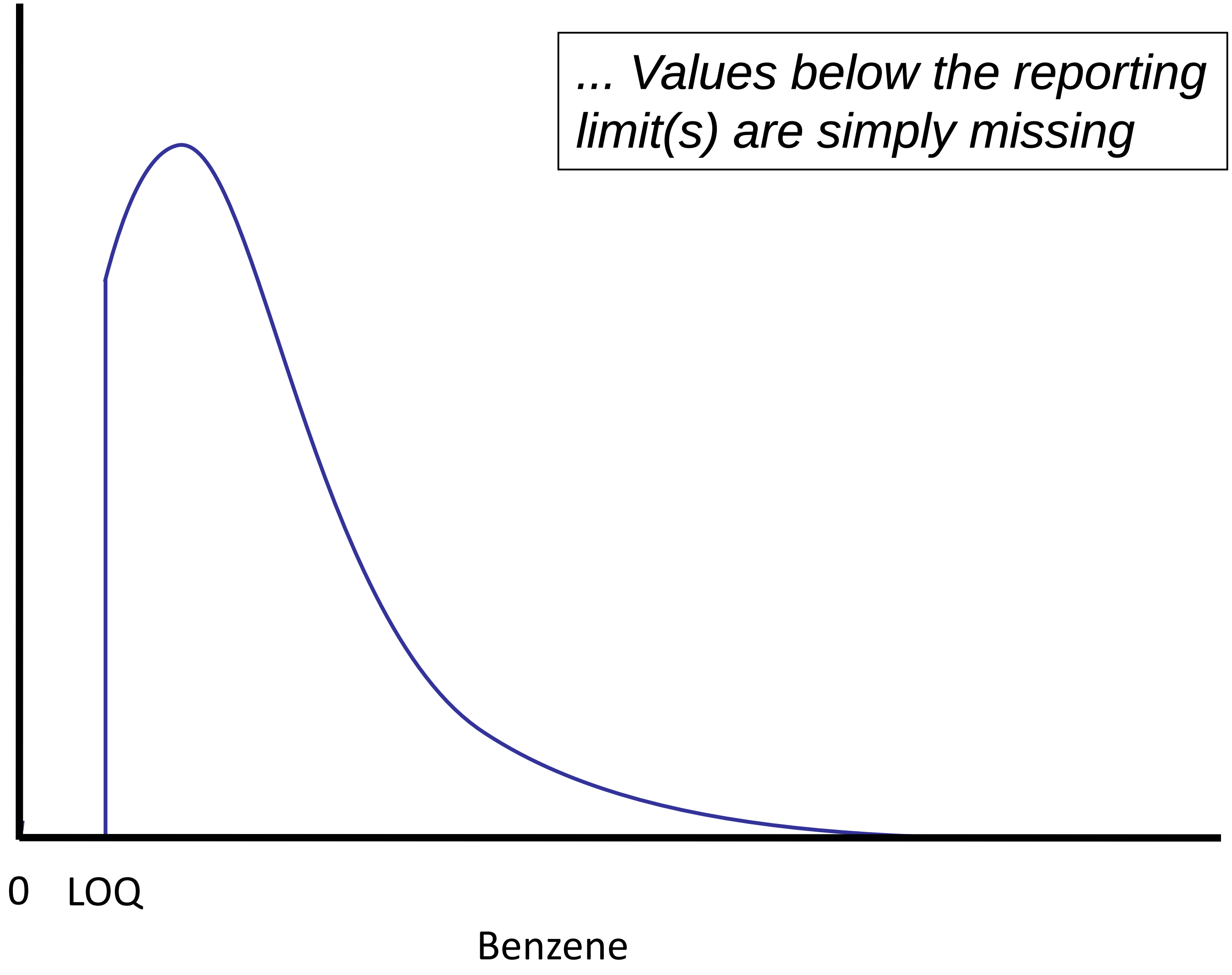
... Bearing in mind that we
want to relate toxic compounds
To fuel properties,
we can think about two (or more)
Dimensions ...

... These are simulated data, but
in terms of our analysis, we can
think of this plot as ln(toxic) vs.
Ethanol or another fuel property

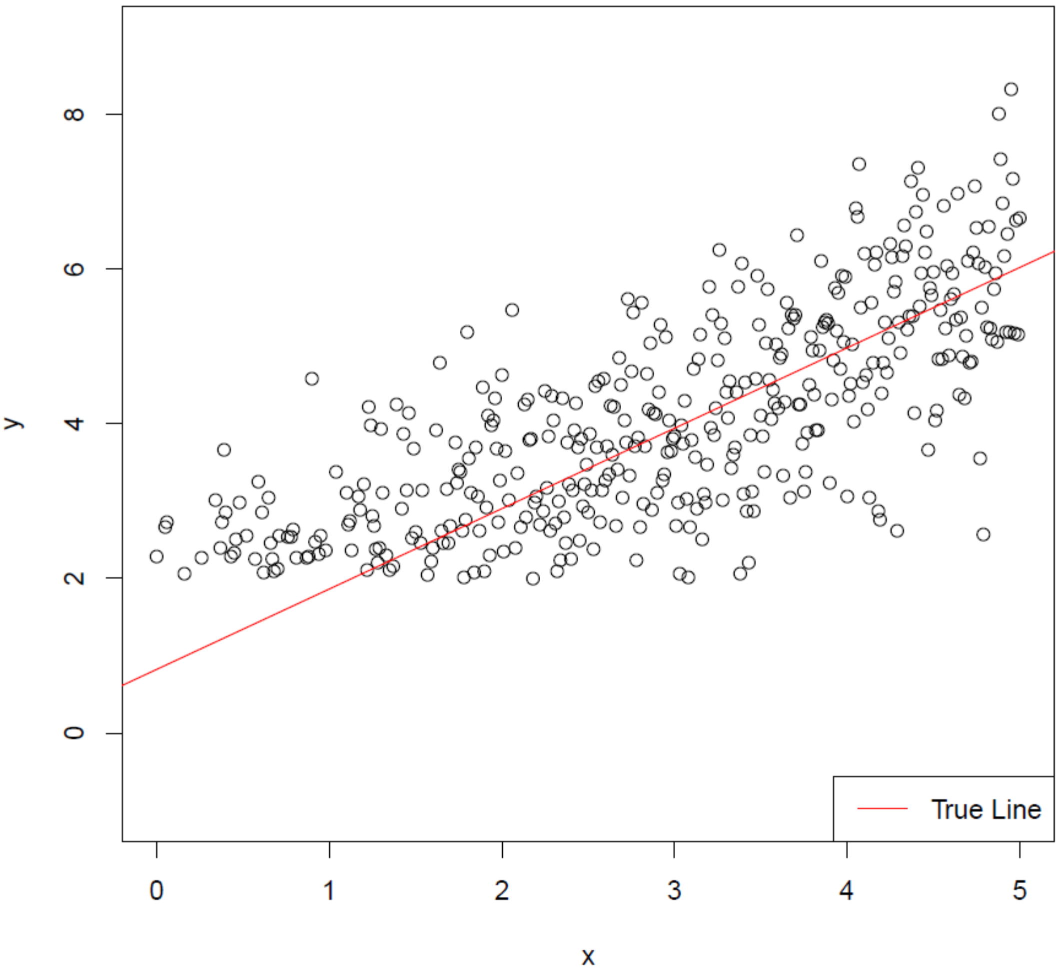


- Line: Y=1+x+error.   error ~ N(0,1)

Censored Distribution



What About Non-Detects?

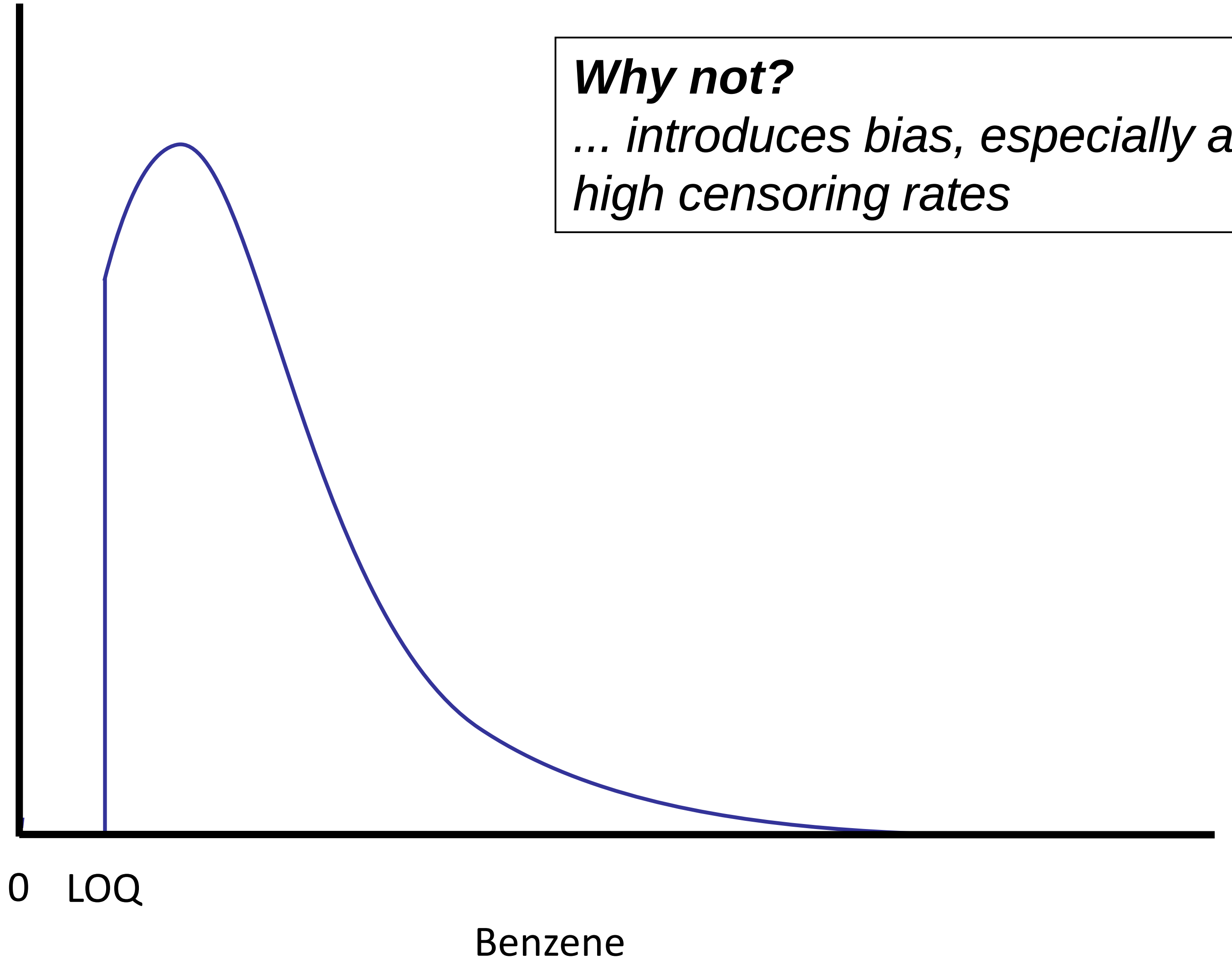


All data points < 2LOQ have been deleted

What to Do?

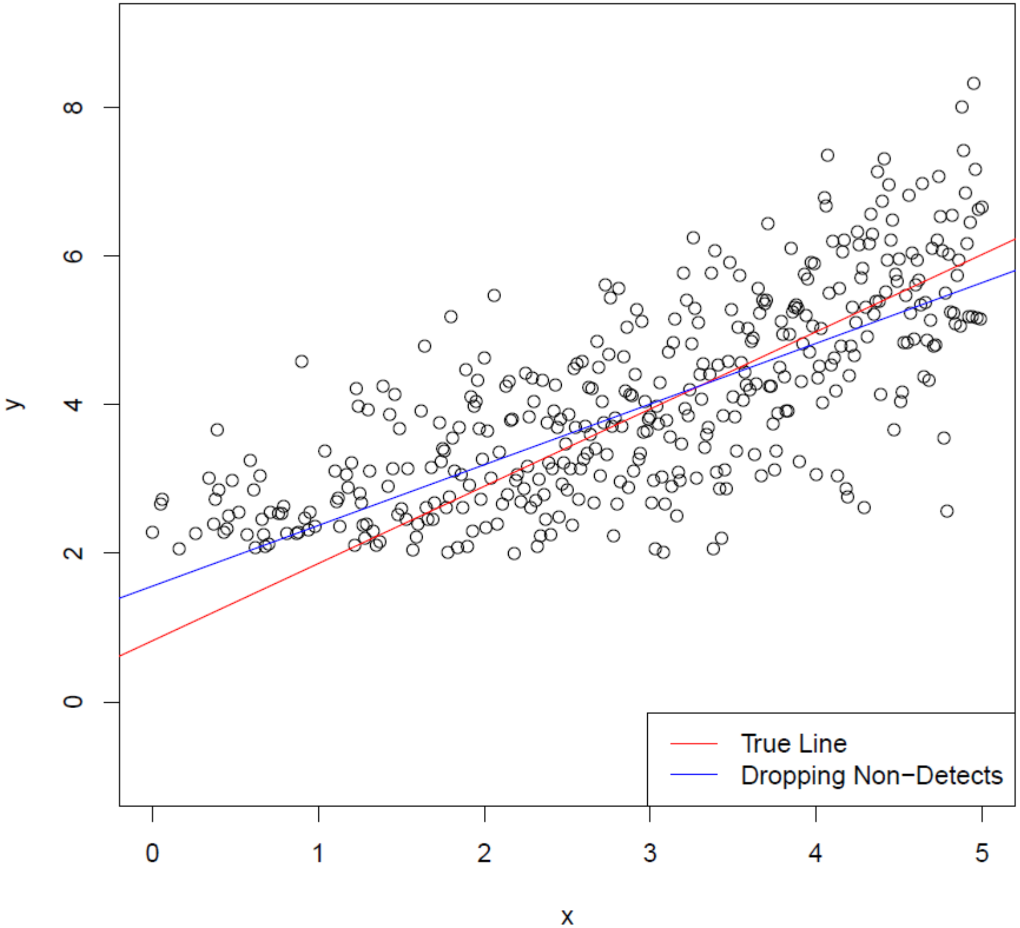
- Variety of approaches developed to address censoring
  - Analyze without the censored data
  - Assign censored values to zero
  - Substitute the limit of detection (LOD)
  - Substitute half the limit of detection (LOD/2)
  - assign random numbers between 0 and LOD
  - Statistical imputation
    - By regression
    - By “maximum likelihood estimation”
    - Other?

Analyze without the censored data



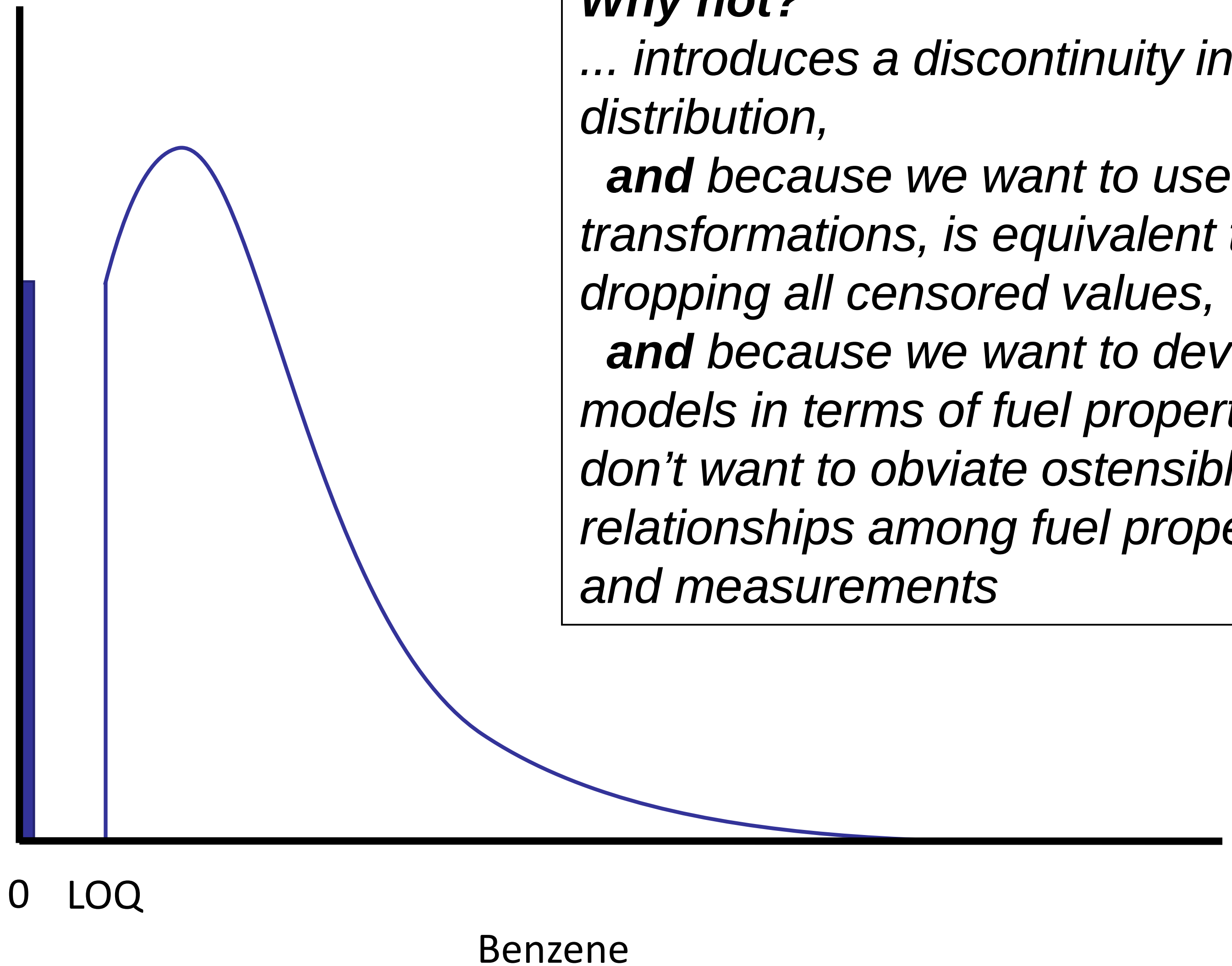
Why not?
... introduces bias, especially at high censoring rates

Why we don't want to just ignore missing points



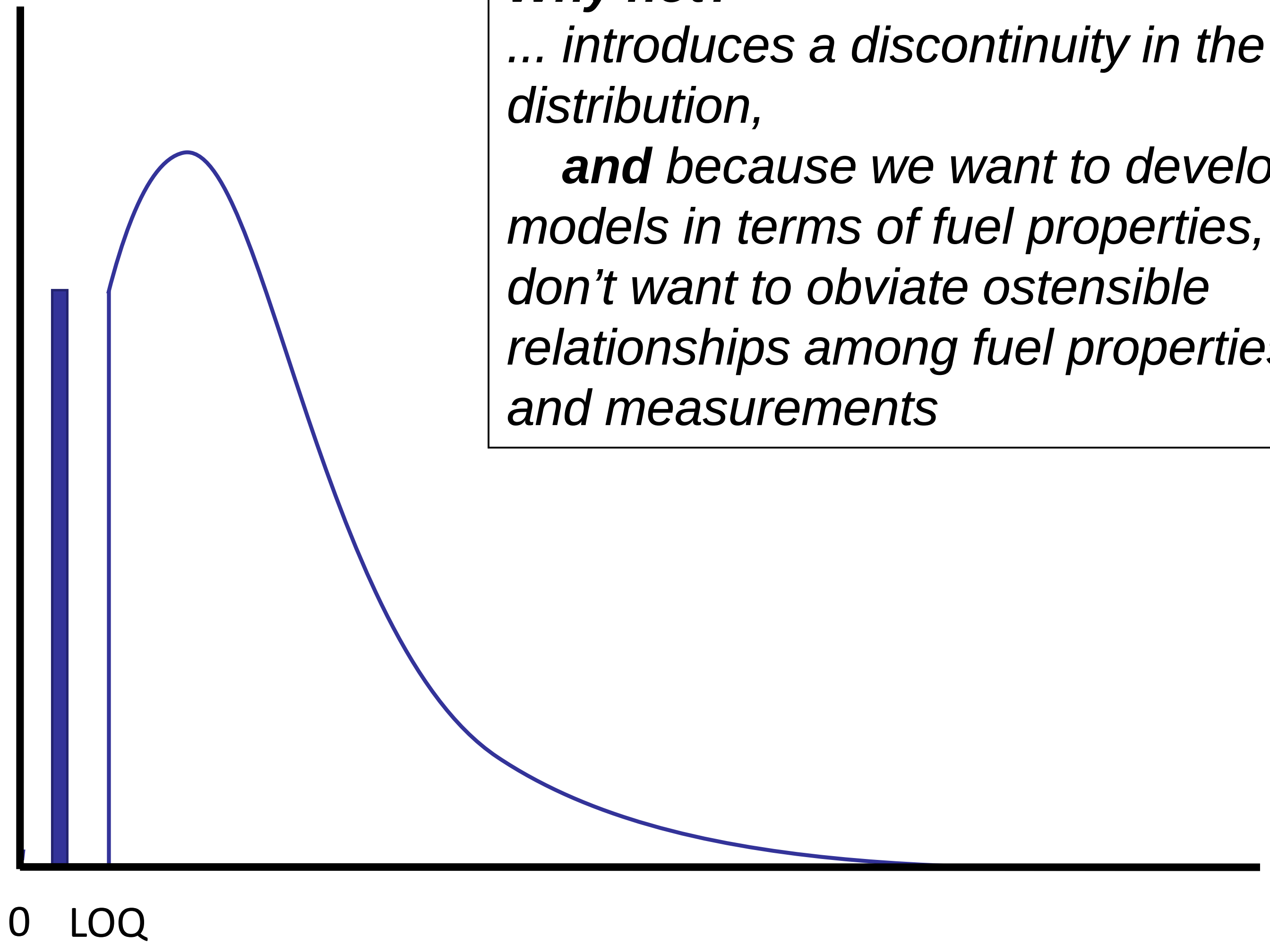
When you fit a line to only the points above the LOQ, you get biased estimates

Assign censored values to 0



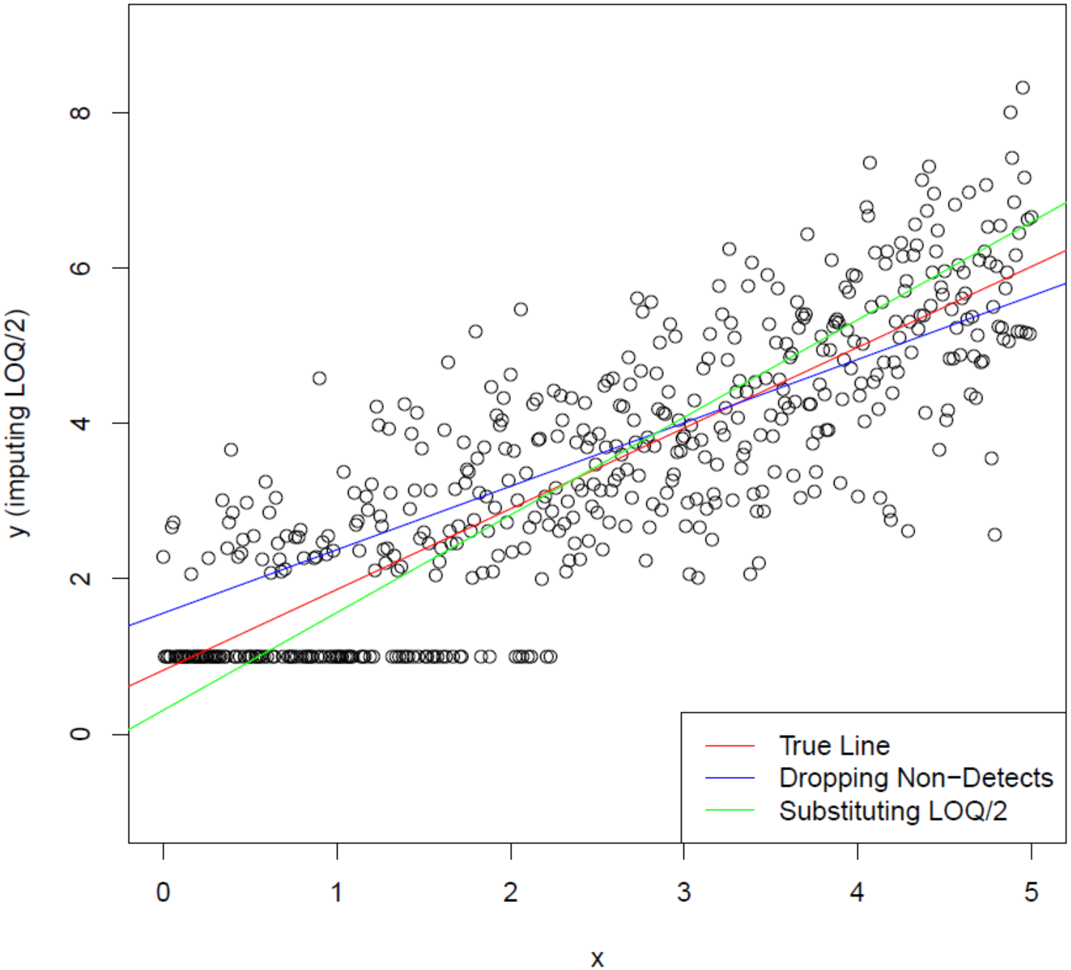
Why not?
... introduces a discontinuity in the distribution,
and *because we want to use log transformations, is equivalent to dropping all censored values,*
and *because we want to develop models in terms of fuel properties, we don't want to obviate ostensible relationships among fuel properties and measurements*

Assign censored values to LOQ/2



Why not?
... introduces a discontinuity in the distribution,
__and__ because we want to develop models in terms of fuel properties, we don't want to obviate ostensible relationships among fuel properties and measurements

What About Just Substituting LOQ/2?



Here LOQ/2 filled in for all the missing points

What About Common Statistical Methods

- “Imputation”
  - Estimate what is not there based on what is there
  - and interrelationships within data
- “Maximum likelihood”
  - Comes in different flavors
  - Estimate where the missing data is “most likely” to be
  - Estimate what unbiased model parameters would “most likely” be
- We (Adam Sales) experimented with several approaches, and
  - We are leaning away from using them
    - They are not appropriate with high censoring rates,
    - They don’t attempt to reconstruct the underlying processes
    - They probably won’t give a substantial improvement over substituting LOQ/2

Another Approach
Estimated Dependent Variable Model (EDV)

- We are uncertain about measurements on the low side of the distribution
  - Are there any emissions from the vehicles?
  - Or just noise?
- Laboratory measurements confounded by
  - contamination from measurement media
    - For particular compounds
  - background contamination
  - fraction of measurement attributable to tailpipe emissions not directly known
    - But may be estimated
- We need additional data
    - Raw uncensored measured values
    - Measurements of media contamination
    - Measurements of background

Step 1: Correct for background and media contamination

First, we assume that toxics measurements are related to fuel properties

$$Y_i = \beta_0 + \beta_1 \cdot \mathrm{etOH} + \beta_2 \cdot \mathrm{ARO} + \beta_3 \cdot T50 + \beta_4 \cdot T90$$

Second, we assume that the true (and unknown) tailpipe toxics measurements are confounded by media (*k*) and background contamination (*b*)

$$\tilde{Y}_i = Y_i + \bar{k}_i + b_i$$

But because both *k* and *b* have been measured, we can take a reasonable shot at estimating the "true" values

$$\hat{Y}_i = \tilde{Y}_i - \bar{k}_i - b_i$$

Step 2: Estimating Variances

- Random error $\left(\hat{\sigma}_{\varepsilon}^{2}\right)$
  - assumptions:
    - Constant over time
    - Not correlated with fuel properties
    - Not serially auto-correlated
- Media contamination $\left(\hat{\sigma}_{k,i}^{2}\right)$
  - Assumptions
    - varies over time
    - Not correlated with fuel properties
    - Not correlated with random error

Estimating the variance of media Contamination

- Option 1
  - Estimate as 5-day moving average of media blanks
  - Previously used to estimate LOQ

$$\hat{\sigma}_{k,i}^2 = \mathrm{Var}\{k_{i-5}, k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}\}$$

- Option 2
  - Estimate as variance of cartridge batches
    - Followup on suggestion (from Dick Gunst)
    - Prelim diagnostics (by Sales) suggest that batch matters
  - Additional data needed (?)

Estimating random error

- Fit an initial model
  - Toxic in terms of fuel properties
  - Obtain residuals (r_i)
  - Re-estimate random error
    - While accounting for variance of media contamination

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{\sum_i r_i^2 - \sum_i \hat{\sigma}_{k,i}^2 + tr\left(\left(\mathbf{X'X}\right)^{-1}\mathbf{X'}diag(\hat{\sigma}_{k,i}^2)\mathbf{X}\right)}{n-p-1}$$

Step 3: Calculate “Variance-based” Weights

- Using the two variances
  - Variance of the media contamination
    - Multiplied by 4.0
      - Enters into picture four times
        - Applies to both bag and media measurements
        - For both primary and secondary cartridges
  - Random error

$$w_i = \frac{1}{\sqrt{4\hat{\sigma}_{k,i}^2 + \hat{\sigma}_{\varepsilon}^2}}$$

Step 4: Generate Final Model

- Estimate final coefficients for fuel effects
- Apply weights w_i to all measurements
- Use “weighted least squares” (WLS)
  - Classic technique to “stabilize variance”
  - Applies “uncertainty penalty” based on media-contamination variance
    - Measurements with high variability in media contamination <u>downweighted</u>
    - Relative to measurements with low variability in media contamination
    - May increase uncertainty in predicted fuel effects